Proximity-based *cis*-regulatory Module Detection using Constraint Programming for Itemset Mining

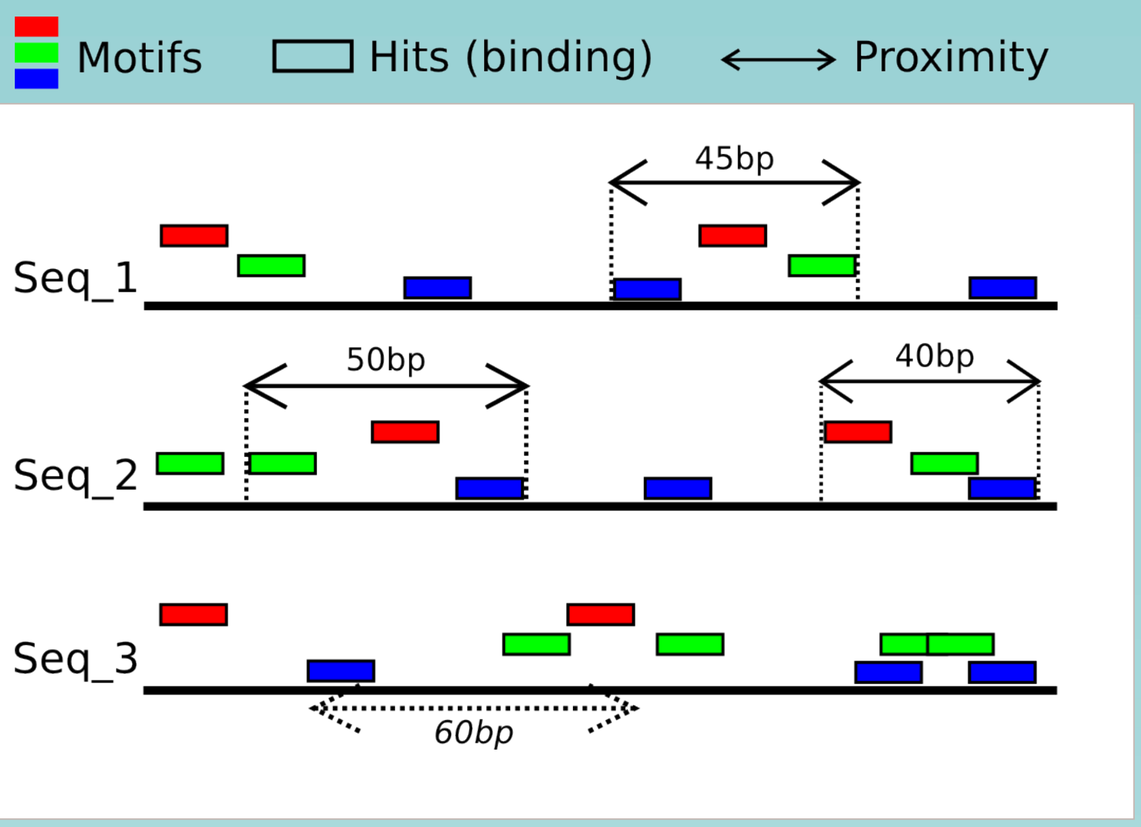
Tias Guns – Hong Sun – Siegfried Nijssen – Aminael Sanchez-Rodriguez – Luc De Raedt – Kathleen Marchal

Problem setting

Single motif discovery tools
→ many false positives

use knowledge across
multiple sequences

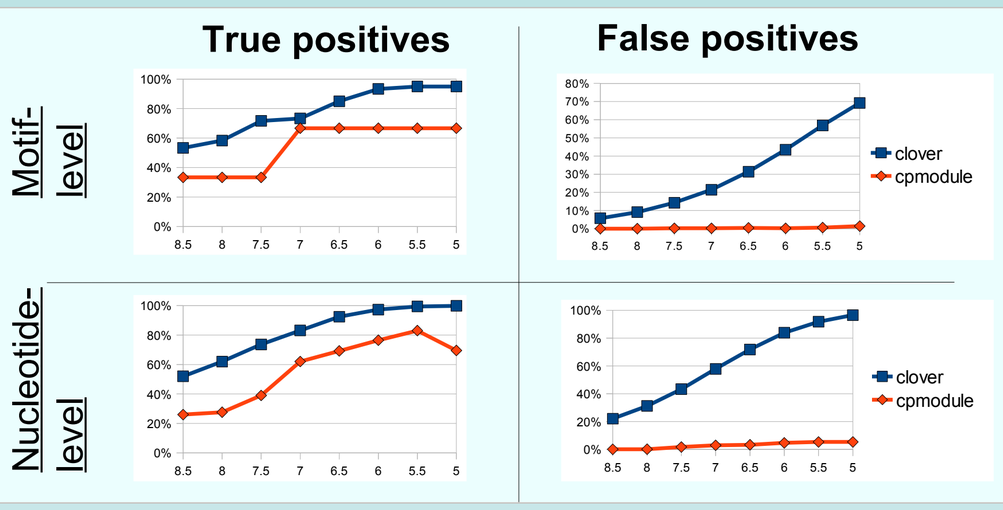
Given: PWMs of motifs, target genomic sequences and background sequences
Find: CRMs involving the same transcription factors in multiple sequences

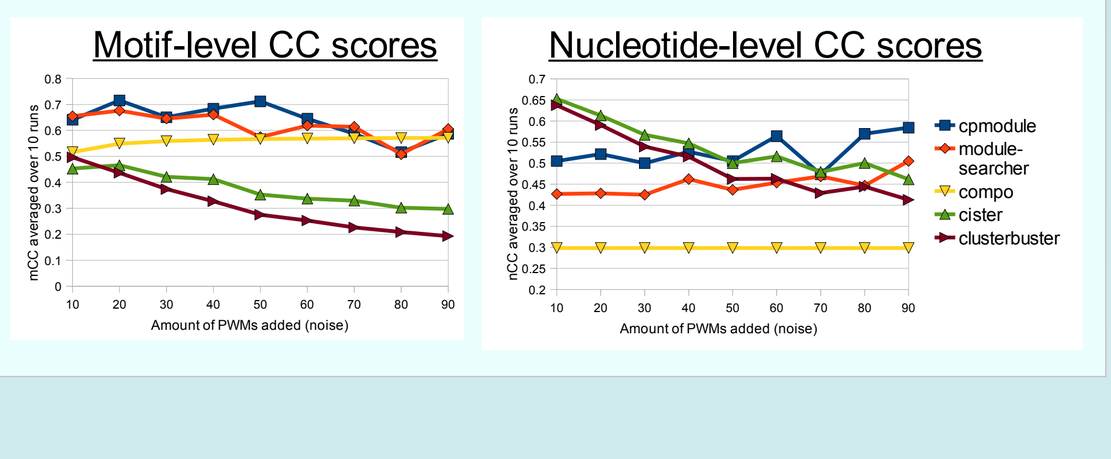


Constraint Programming

General methodology for handling constraint satisfaction problems.

Constraints, on a set of motifs:
- Proximity: only the motifs' hits that bind in each others proximity are considered,
- Coverage: a sequence is covered if the motifs satisfy the *proximity* constraint on it,
- Frequency: the motifs have to *cover* a sufficient number of sequences,
- Redundancy: if two related motif-sets have the same *frequency*, remove the smaller one.





Conclusions

- Principled and flexible approach.
- Significant reduction of false positives, without reduction of true positives.
- Competitive or better predictive performance.

Future work:
- Add more constraints (overlap, priors, ...)
- Other data sources (ChipSeq, ...)

KATHOLIEKE UNIVERSITEIT LEUVEN
DTAI

CP4IM http://www.cs.kuleuven.ac.be/~dtai/CP4IM/ CP4IM